

# Linear-time Computation of Minimal Absent Words Using Suffix Array

Carl Barton<sup>1</sup>, Alice Heliou<sup>2,3</sup>, Laurent Mouchard<sup>4</sup>, and Solon P. Pissis<sup>1</sup>

<sup>1</sup> Department of Informatics, King's College London, London, UK  
`{carl.barton,solon.pissis}@kcl.ac.uk`

<sup>2</sup> Inria Saclay-Île de France, AMIB, Bâtiment Alan Turing, France

<sup>3</sup> Laboratoire d'Informatique de l'École Polytechnique (LIX), CNRS UMR 7161, France

`alice.heliou@polytechnique.org`

<sup>4</sup> University of Rouen, LITIS EA 4108, TIBS, Rouen, France  
`laurent.mouchard@univ-rouen.fr`

**Abstract.** An *absent word* of a word  $y$  of length  $n$  is a word that does not occur in  $y$ . It is a *minimal absent word* if all its proper factors occur in  $y$ . Minimal absent words have been computed in genomes of organisms from all domains of life; their computation provides a fast alternative for measuring approximation in sequence comparison. There exists an  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for computing all minimal absent words on a fixed-sized alphabet based on the construction of suffix automata (Crochemore et al., 1998). No implementation of this algorithm is publicly available. There also exists an  $\mathcal{O}(n^2)$ -time and  $\mathcal{O}(n)$ -space algorithm for the same problem based on the construction of suffix arrays (Pinho et al., 2009). An implementation of this algorithm was also provided by the authors and is currently the fastest available. In this article, we bridge this unpleasant gap by presenting an  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for computing all minimal absent words based on the construction of suffix arrays. Experimental results using real and synthetic data show that the respective implementation outperforms the one by Pinho et al.

## 1 Introduction

Sequence comparison is an important step in many important tasks in bioinformatics. It is used in many applications; from phylogenies reconstruction to the reconstruction of genomes. Traditional techniques for measuring approximation in sequence comparison are based on the notions of distance or of similarity between sequences; and these are computed through sequence alignment techniques. An issue with using alignment techniques is that they are computationally expensive: they require quadratic time in the length of the sequences. Moreover, in molecular taxonomy and phylogeny, for instance, whole-genome alignment proves both computationally expensive and hardly significant. These observations have led to increased research into *alignment free* techniques for sequence comparison. A number of alignment free techniques have been proposed:

in [10], a method based on the computation of the shortest unique factors of each sequence is proposed; other approaches estimate the number of mismatches per site based on the length of exact matches between pairs of sequences [5].

Thus standard notions are gradually being complemented (or even supplanted) by other measures that refer, implicitly or explicitly, to the composition of sequences in terms of their constituent patterns. One such measure is the notion of words absent in a sequence. A word is an *absent word* of some sequence if it does not occur in the sequence. These words represent a type of *negative information*: information about what does not occur in the sequence. Noting the words which do occur in one sequence but do not occur in another can be used to detect mutations or other biologically significant events.

Given a sequence of length  $n$ , the number of absent words of length at most  $n$  can be exponential in  $n$ , meaning that using all the absent words for sequence comparison is more expensive than alignments. However, the number of certain subsets of absent words is only linear in  $n$ . An absent word of a sequence is a *shortest absent word* if all words shorter than it do occur in the sequence. An  $\mathcal{O}(mn)$ -time algorithm for computing shortest absent words was presented in [9], where  $m$  is a user-specified threshold on the length of the shortest absent words. This was later improved by [11], who presented an  $\mathcal{O}(n \log \log n)$ -time algorithm for the same problem. This has been further improved and an  $\mathcal{O}(n)$ -time algorithm was presented in [16].

A *minimal absent word* of a sequence is an absent word whose proper factors all occur in the sequence. Notice that minimal absent words are a superset of shortest absent words [15]. An upper bound on the number of minimal absent words is  $\mathcal{O}(\sigma n)$  [4,13], where  $\sigma$  is the size of the alphabet. This suggests that it may be possible to compare sequences in time proportional to their lengths, for a fixed-sized alphabet, instead of proportional to the product of their lengths [10].

Recently, there has been a number of biological studies on the significance of absent words. The most comprehensive study on the significance of absent words is probably [1]; in this, the authors suggest that the deficit of certain subsets of absent words in vertebrates may be explained by the hypermutability of the genome. It was later found in [8] that the compositional biases observed in [1] for vertebrates are not uniform throughout different sets of minimal absent words. Moreover, the analyses in [8] support the hypothesis of the inheritance of minimal absent words through a common ancestor, in addition to lineage-specific inheritance, only in vertebrates. In [7], the minimal absent words in four human genomes were computed, and it was shown that, as expected, intra-species variations in minimal absent words were lower than inter-species variations. Minimal absent words have also been used for phylogenies reconstruction [3].

From an algorithmic perspective, an  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for computing all minimal absent words on a fixed-sized alphabet based on the construction of suffix automata was presented in [4]. An alternative  $\mathcal{O}(n)$ -time solution for finding minimal absent words of length at most  $\ell$ , such that  $\ell = \mathcal{O}(1)$ , based on the construction of tries of bounded-length factors was presented in [3]. A drawback of these approaches, in practical terms, is that the construction of

suffix automata (or of tries) may have a large memory footprint. Due to this, an important problem is to be able to compute the minimal absent words of a sequence without the use of data structures such as the suffix automaton. To this end, the computation of minimal absent words based on the construction of suffix arrays was considered in [15]; although fast in practice, the worst-case runtime of this algorithm is  $\mathcal{O}(n^2)$ . Alternatively, one could make use of the succinct representations of the bidirectional BWT, recently presented in [2], to compute all minimal absent words in time  $\mathcal{O}(n)$ . However, an implementation of these representations was not made available by the authors; and it is also rather unlikely that such an implementation will outperform an  $\mathcal{O}(n)$ -time algorithm based on the construction of suffix arrays.

*Our Contribution:* In this article, we bridge this unpleasant gap by presenting the first  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for computing all minimal absent words of a sequence of length  $n$  based on the construction of suffix arrays. The respective implementation is also provided and shown to be more efficient than existing tools.

## 2 Definitions and Notation

To provide an overview of our result and algorithm, we begin with a few definitions. Let  $y = y[0]y[1] \dots y[n-1]$  be a *word* of *length*  $n = |y|$  over a finite ordered *alphabet*  $\Sigma$  of size  $\sigma = |\Sigma| = \mathcal{O}(1)$ . We denote by  $y[i \dots j] = y[i] \dots y[j]$  the *factor* of  $y$  that starts at position  $i$  and ends at position  $j$  and by  $\varepsilon$  the *empty word*, word of length 0. We recall that a *prefix* of  $y$  is a factor that starts at position 0 ( $y[0 \dots j]$ ) and a *suffix* is a factor that ends at position  $n-1$  ( $y[i \dots n-1]$ ), and that a factor of  $y$  is a *proper factor* if it is not the empty word or  $y$  itself.

Let  $x$  be a word of length  $0 < m \leq n$ . We say that there exists an *occurrence* of  $x$  in  $y$ , or, more simply, that  $x$  *occurs in*  $y$ , when  $x$  is a factor of  $y$ . Every occurrence of  $x$  can be characterised by a starting position in  $y$ . Thus we say that  $x$  occurs at the *starting position*  $i$  in  $y$  when  $x = y[i \dots i+m-1]$ . Opposingly, we say that the word  $x$  is an *absent word* of  $y$  if it does not occur in  $y$ . The absent word  $x$ ,  $m \geq 2$ , of  $y$  is *minimal* if and only if all its proper factors occur in  $y$ .

We denote by **SA** the *suffix array* of  $y$ , that is the array of length  $n$  of the starting positions of all sorted suffixes of  $y$ , i.e. for all  $1 \leq r < n-1$ , we have  $y[\text{SA}[r-1] \dots n-1] < y[\text{SA}[r] \dots n-1]$  [12]. Let  $\text{lcp}(r, s)$  denote the length of the longest common prefix of the words  $y[\text{SA}[r] \dots n-1]$  and  $y[\text{SA}[s] \dots n-1]$ , for all  $0 \leq r, s < n-1$ , and 0 otherwise. We denote by **LCP** the *longest common prefix* array of  $y$  defined by  $\text{LCP}[r] = \text{lcp}(r-1, r)$ , for all  $1 < r < n-1$ , and  $\text{LCP}[0] = 0$ . The inverse **iSA** of the array **SA** is defined by  $\text{iSA}[\text{SA}[r]] = r$ , for all  $0 \leq r < n-1$ . **SA** [14], **iSA**, and **LCP** [6] of  $y$  can be computed in time and space  $\mathcal{O}(n)$ .

In this article, we consider the following problem.

MINIMALABSENTWORDS

**Input:** a word  $y$  on  $\Sigma$  of length  $n$

**Output:** all tuples  $\langle a, (i, j) \rangle$ , such that word  $x$ , defined by  $x[0] = a, a \in \Sigma$ , and  $x[1..m-1] = y[i..j]$ ,  $m \geq 2$ , is a minimal absent word of  $y$

### 3 Algorithm MAW

In this section, we present algorithm MAW, an  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for finding all minimal absent words in a word of length  $n$  using arrays SA and LCP. We first explain how we can characterise the minimal absent words; then we introduce how their computation can be done efficiently by using arrays SA and LCP. Finally, we present in detail the two main steps of the algorithm.

A minimal absent word  $x[0..m-1]$  of a word  $y[0..n-1]$  is an absent word whose proper factors all occur in  $y$ . Among them,  $x_1 = x[1..m-1]$  and  $x_2 = x[1..m-2]$  occur in  $y$ ; we will focus on these two factors to characterise the minimal absent words. To do so, we will consider each occurrence of  $x_1$  and  $x_2$ , and construct the sets of letters that occur just before:

$$B(x_1) = \{y[j-1] : j \text{ is the starting position of an occurrence of } x_1\}$$

$$B(x_2) = \{y[j-1] : j \text{ is the starting position of an occurrence of } x_2\}$$

**Lemma 1.** *Let  $x$  and  $y$  be two words. Then  $x$  is a minimal absent word of  $y$  if and only if  $x[0]$  is an element of  $B(x_2)$  and not of  $B(x_1)$ , with  $x_1 = x[1..m-1]$  and  $x_2 = x[1..m-2]$ .*

*Proof.* ( $\Rightarrow$ ) Let  $x_1$  be a factor of  $y$ ,  $x_2$  be the longest proper prefix of  $x_1$ , and  $B(x_1)$  and  $B(x_2)$  the sets defined above. Further let  $p$  be a letter that is in  $B(x_2)$  but not in  $B(x_1)$ . Then, there exists a starting position  $j$  of an occurrence of  $x_2$  such that  $y[j-1] = p$ , so the word  $px_2$  occurs at position  $j-1$  in  $y$ .  $p$  is not in  $B(x_1)$  so  $px_1$  does not occur in  $x$  and is therefore an absent word of  $y$ .  $x_1$  and  $px_2$  are factors of  $y$ , so all the proper factors of  $px_1$  occur in  $y$ , thus  $px_1$  is a minimal absent word of  $y$ .

( $\Leftarrow$ ) Let  $x[0..m-1]$  be a minimal absent word of  $y$ . Its longest proper prefix  $x[0..m-2] = x[0]x_2$  occurs in  $y$ , so  $x[0]$  is in  $B(x_2)$ . Its longest proper suffix,  $x_1$  occurs as well in  $y$ , but  $x = x[0]x_1$  is an absent word of  $y$  so it does not occur in  $y$  and  $x[0]$  is not in  $B(x_1)$ .  $\square$

**Lemma 2.** *Let  $x$  be a minimal absent word of length  $m$  of word  $y$  of length  $n$ . Then there exists an integer  $i \in [0 : n-1]$  such that  $y[\text{SA}[i].. \text{SA}[i] + \text{LCP}[i]] = x_1$  or  $y[\text{SA}[i].. \text{SA}[i] + \text{LCP}[i+1]] = x_1$ , where  $x_1 = x[1..m-1]$ .*

*Proof.* Let  $j$  be the starting position of an occurrence of  $x[0..m-2]$  in  $y$  and  $k$  the starting position of an occurrence of  $x_1$  in  $y$ . The suffixes  $y[j+1..n-1]$  and  $y[k..n-1]$  share  $x_2 = x[1..m-2]$  as a common prefix. As  $x$  is an absent word of  $y$ , this common prefix cannot be extended so  $x_2$  is the longest common prefix of those suffixes. By using iSA, the inverse suffix array, we have  $\text{lcp}(\text{iSA}[j+1..n], \text{iSA}[k..n]) = m-2$ .

$1], \text{iSA}[k]) = m - 2$ . Let us also note  $s_k = \text{iSA}[k]$  and  $s_{j+1} = \text{iSA}[j + 1]$ . We then have two possibilities:

- if  $s_k > s_{j+1}$ : for all  $s$  in  $[s_{j+1} + 1 : s_k]$ , we have  $\text{LCP}[s] \geq m - 2$ , with equality holding for at least one position. Let us define  $i = \max\{s \in [s_{j+1} : s_k] : \text{LCP}[s] = m - 2\}$ , the maximality of  $i$  implies that  $i = s_k$  or  $\text{lcp}(i, s_k) > m - 2$  and thus, in both cases  $y[\text{SA}[i] \dots \text{SA}[i] + \text{LCP}[i]] = x_1$ .
- if  $s_{j+1} > s_k$ : for all  $s$  in  $[s_k + 1 : s_{j+1}]$ , we have  $\text{LCP}[s] \geq m - 2$ , with equality holding for at least one position. Let us define  $i = \min\{s \in [s_k : s_{j+1}] : \text{LCP}[s + 1] = m - 2\}$ , the minimality of  $i$  implies  $i = s_k$  or  $\text{lcp}(s_k, i) > m - 2$  and thus, in both cases  $y[\text{SA}[i] \dots \text{SA}[i] + \text{LCP}[i + 1]] = x_1$ .

For an illustration inspect Fig. 1. □

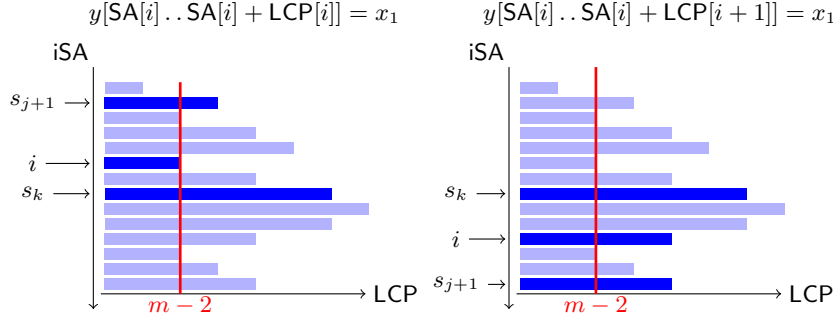


Fig. 1: Illustration of Lemma 2

By Lemma 2, we can compute all minimal absent words of  $y$  by examining only the factors  $S_{2i} = y[\text{SA}[i] \dots \text{SA}[i] + \text{LCP}[i]]$  and  $S_{2i+1} = y[\text{SA}[i] \dots \text{SA}[i] + \text{LCP}[i + 1]]$ , for all  $i$  in  $[0 : n - 1]$ . We just need to construct the sets  $B_1(S_{2i})$ ,  $B_2(S_{2i})$  and  $B_1(S_{2i+1})$ ,  $B_2(S_{2i+1})$ , where  $B_1(S_j)$  (resp.  $B_2(S_j)$ ) is the set of letters that immediately precede an occurrence of the factor  $S_j$  (resp. the longest proper prefix of  $S_j$ ), for all  $j$  in  $[0 : 2n - 1]$ . Then, by Lemma 1, the symmetric difference between  $B_{S_j,1}$  and  $B_{S_j,2}$ , for all  $j$  in  $[0 : 2n - 1]$ , gives us all the minimal absent words of  $y$ .

Thus the important computational step is to compute these sets of letters efficiently. To do so, we visit twice arrays **SA** and **LCP** using another array denoted by **Before** (resp. **BeforeLCP**) to store set  $B_1(S_j)$  (resp.  $B_2(S_j)$ ), for all  $j$  in  $[0 : 2n - 1]$ . Both arrays **Before** and **BeforeLCP** consist of  $2n$  elements, where each element is a bit vector of length  $\sigma$ , the size of the alphabet, corresponding to one bit per alphabet letter. While iterating over arrays **SA** and **LCP**, we maintain another array denoted by **Interval**, such that, at the end of each iteration  $i$ , the  $\ell^{\text{th}}$  element of **Interval** stores the set of letters we have encountered before the

prefix of length  $\ell$  of  $y[\text{SA}[i] \dots n - 1]$ . Array **Interval** consists of  $\max_{i \in [0:n-1]} \text{LCP}[i] + 1$  elements, where each element is a bit vector of length  $\sigma$ .

During the first pass, we visit arrays **SA** and **LCP** from top to bottom. For each  $i \in [0 : n - 1]$ , we store in positions  $2i$  and  $2i + 1$  of **Before** (resp. **BeforeLCP**) the set of letters that immediately precede occurrences of  $S_{2i}$  and  $S_{2i+1}$  (resp. their longest proper prefixes) whose starting positions appear before position  $i$  in **SA**. During the second pass, we go bottom up to complete the sets, which are already stored, with the letters preceding the occurrences whose starting positions appear after position  $i$  in **SA**. In order to be efficient, we will maintain a stack structure, denoted by **LifoLCP**, to store the **LCP** values of the factors that are prefixes of the one we are currently visiting.

**Function** *Top-Down-Pass* ( $y, n, \text{SA}, \text{LCP}, \text{Before}, \text{BeforeLCP}, \sigma$ )

```

Interval[0 .. max_{i \in [0:n-1]} LCP[i]][0 .. \sigma - 1] \leftarrow 0;
LifoLCP.push(0);
foreach  $i \in [0 : n - 1]$  do
  if  $i > 0$  and  $\text{LCP}[i] < \text{LCP}[i - 1]$  then
    while LifoLCP.top() > LCP[i] do
      proxa \leftarrow LifoLCP.pop();
      Interval[proxa][0 .. \sigma - 1] \leftarrow 0;
    if LifoLCP.top() < LCP[i] then
      Interval[LCP[i]] \leftarrow Interval[proxa];
      Before[2i - 1] \leftarrow Interval[proxa];
      BeforeLCP[2i - 1] \leftarrow Interval[LCP[i]];
  if  $\text{SA}[i] > 0$  then
     $u \leftarrow y[\text{SA}[i] - 1]$ ; value \leftarrow LifoLCP.top();
    while Interval[value][u] = 0 do
      Interval[value][u] \leftarrow 1; value \leftarrow LifoLCP.next();
    Interval[LCP[i]][u] \leftarrow 1;
    Before[2i][u] \leftarrow 1; Before[2i + 1][u] \leftarrow 1;
    BeforeLCP[2i][u] \leftarrow 1; BeforeLCP[2i + 1][u] \leftarrow 1;
  if  $i > 0$  and  $\text{LCP}[i] > 0$  and  $\text{SA}[i - 1] > 0$  then
     $v \leftarrow y[\text{SA}[i - 1] - 1]$ ;
    Interval[LCP[i]][v] \leftarrow 1;
    BeforeLCP[2i] \leftarrow Interval[LCP[i]];
  if LifoLCP.top() \neq LCP[i] then LifoLCP.push(LCP[i])

```

### 3.1 Top-down Pass

Each iteration of the top-down pass consists of two steps. In the first step, we visit **LifoLCP** from the top and for each **LCP** value read we set to zero the

(a)

$j$	Before	BeforeLCP
0	00	00
1	00	00
2	10	10
3	10	10
4	01	11
5	11	11
6	01	11
7	11	11
8	01	11
9	01	01
10	10	11
11	10	10
12	10	10
13	10	11
14	10	11

(b)

$i$	LCP	SA	Factor	Interval[0]	Interval[1]	Interval[2]	Interval[3]	Interval[4]
0	0	0	A	00	00	00	00	00
			A A					
1	1	1	A B	10	10	00	00	00
			A B A B A					
2	4	3	A B A B B	11	11	00	00	11
			A B A					
3	2	5	A B B	11	11	11	00	00
			A					
4	0	7	B	11	00	00	00	00
			B					
5	1	2	B A	11	11	00	00	00
			B A B A					
6	3	4	B A B B	11	11	00	10	00
			B A					
7	1	6	B B	11	11	00	00	00

Table 1: (a) Arrays **Before** and **BeforeLCP** obtained after the top-down pass for word  $y = \text{AABABABB}$ ; (b) Elements of array **Interval** at the end of each iteration of the top-down pass. Factors  $S_j$  are in orange and violet; their longest proper prefixes are in orange only.

corresponding element of **Interval**; then we remove this value from the stack. We stop when we reach a value smaller or equal to  $\text{LCP}[i]$ . We do this as the corresponding factors are not prefixes of  $y[\text{SA}[i] \dots n-1]$ , nor will they be prefixes in the remaining suffixes. We push at most one value onto the stack **LifoLCP** per iteration, so, in total, there are  $n$  times we will set an element of **Interval** to zero. This step requires time and space  $\mathcal{O}(n\sigma)$ .

For the second step, we update the elements that correspond to factors in the suffix array with an LCP value less than  $\text{LCP}[i]$ . To do so, we visit the stack **LifoLCP** top-down and, for each LCP value read, we add the letter  $y[\text{SA}[i] - 1]$  to the corresponding element of **Interval** until we reach a value whose element already contains it. This ensures that, for each value read, the corresponding element of **Interval** has no more than  $\sigma$  letters added. As we consider at most  $n$  values, this step requires time and space  $\mathcal{O}(n\sigma)$ . For an example, see Table 1.

### 3.2 Bottom-up Pass

Intuitively, the idea behind the bottom-up pass is the same as in the top-down pass except that in this instance, as we start from the bottom, the suffix  $y[\text{SA}[i] \dots n-1]$  can share more than its prefix of length  $\text{LCP}[i]$  with the previous suffixes in **SA**. Therefore we may need the elements of **Interval** that correspond to factors with an LCP value greater than  $\text{LCP}[i]$  to correctly compute the arrays **Before** and **BeforeLCP**. To achieve this, we maintain another stack **LifoRem** to copy the values from **LifoLCP** that are greater than  $\text{LCP}[i]$ . This extra stack allows us to keep in **LifoLCP** only values that are smaller or equal to  $\text{LCP}[i]$  without losing the additional information we require to correctly compute **Before** and

BeforeLCP. At the end of the iteration, we will set to zero each element corresponding to a value in LifoRem and empty the stack. Thus to set an element of Interval to zero requires two operations more than in the first pass. As we consider at most  $n$  values, this step requires time and space  $\mathcal{O}(n\sigma)$ .

```

Function Bottom-Up-Pass( $n$ , SA, LCP, Before, BeforeLCP,  $\Sigma$ ,  $\sigma$ )
  Interval[0 ..  $\max_{i \in [0:n-1]} \text{LCP}[i][0.. \sigma - 1]$ ]  $\leftarrow$  0;
  LifoLCP.push(0);
  foreach  $i \in [n - 1 : 0]$  do
    proxa  $\leftarrow$  LCP[ $i$ ] + 1; proxb  $\leftarrow$  1;
    if  $i < n - 1$  and LCP[ $i$ ] < LCP[ $i + 1$ ] then
      while LifoLCP.top() > LCP[ $i$ ] do
        proxa  $\leftarrow$  LifoLCP.pop();
        LifoRem.push(proxa);
      if LifoLCP.top() < LCP[ $i$ ] then
        Interval[LCP[ $i$ ]]  $\leftarrow$  Interval[proxa]
    foreach  $k \in \Sigma : \text{Before}[2i][k] = 1$  do
      value  $\leftarrow$  LifoLCP.top();
      while Interval[value][ $k$ ] = 0 do
        Interval[value][ $k$ ]  $\leftarrow$  1; value  $\leftarrow$  LifoLCP.next();
      Interval[LCP[ $i$ ]][ $k$ ]  $\leftarrow$  1;

    BeforeLCP[2 $i$ ]  $\leftarrow$  BeforeLCP[2 $i$ ] bit-or Interval[LCP[ $i$ ]];
    BeforeLCP[2 $i + 1$ ]  $\leftarrow$  BeforeLCP[2 $i + 1$ ] bit-or Interval[LCP[ $i + 1$ ]];
    Before[2 $i + 1$ ]  $\leftarrow$  Before[2 $i + 1$ ] bit-or Interval[proxb];

    proxb  $\leftarrow$  proxa;
    Before[2 $i$ ]  $\leftarrow$  Before[2 $i$ ] bit-or Interval[proxa];
    while LifoRem not empty do
      value  $\leftarrow$  LifoRem.pop(); Interval[value][0 ..  $\sigma - 1$ ]  $\leftarrow$  0;
    if LifoLCP.top()  $\neq$  LCP[ $i$ ] then LifoLCP.push(LCP[ $i$ ])

```

Another difference between the top-down and bottom-up passes is that in order to retain the information computed in the first pass, the second step is performed for each letter in Before[2 $i$ ]. As, for each LCP value read, we still add a letter only if it is not already contained in the corresponding element of Interval, no more than  $\sigma$  letters are added. Thus this step requires time and space  $\mathcal{O}(n\sigma)$ . For an example, see Table 2.

Once we have computed arrays Before and BeforeLCP, we need to compare each element. If there is a symmetric difference, by Lemma 1, we can construct a minimal absent word. For an example, see Table 3. To ensure that we can report the minimal absent words in linear time, we must be able to report each one in constant time. To achieve this, we can represent them as a tuple  $\langle a, (i, j) \rangle$ , where for some word  $x$  of length  $m \geq 2$  that is a minimal absent word of  $y$ , the



(a)

$j$	Before	BeforeLCP
0	11	11
1	00	11
2	11	11
3	10	11
4	01	11
5	11	11
6	01	11
7	11	11
8	11	11
9	01	11
10	10	11
11	10	10
12	10	10
13	10	11
14	10	11

(b)

$i$	LCP	SA	Factor	Interval[0]	Interval[1]	Interval[2]	Interval[3]	Interval[4]
7	1	6	B B	10	10	00	00	00
			B A					
6	3	4	B A B B	10	10	00	10	00
			B A B A					
5	1	2	B A	10	10	00	00	00
			B					
4	0	7	B	11	00	00	00	00
			A					
3	2	5	A B B	11	00	01	00	00
			A B A					
2	4	3	A B A B B	11	00	01	00	01
			A B A B A					
1	1	1	A B	11	11	00	00	00
			A A					
0	0	0	A	11	00	00	00	00

Table 2: (a) Arrays Before and BeforeLCP obtained after the bottom-up pass for word  $y = \text{AABABABB}$ ; (b) Elements of array Interval at the end of each iteration of the bottom-up pass. Factors  $S_j$  are in orange and violet; their longest proper prefixes are in orange only.

following holds:  $x[0] = a$  and  $x[1..m-1] = y[i..j]$ . Lemma 2 ensures us to be exhaustive. Therefore we obtain the following result.

**Theorem 1.** *Algorithm MAW solves problem MINIMALABSENTWORDS in time and space  $\mathcal{O}(n)$ .*

## 4 Experimental Results

We implemented algorithm MAW as a programme to compute all minimal absent words of a given sequence. The programme was implemented in the C programming language and developed under GNU/Linux operating system. It takes as input arguments a file in (Multi)FASTA format and the minimal and maximal length of minimal absent words to be outputted; and then produces a file with all minimal absent words of length within this range as output. The implementation is distributed under the GNU General Public License (GPL), and it is available at <http://github.com/solonas13/maw>, which is set up for maintaining the source code and the man-page documentation. The experiments were conducted on a Desktop PC using one core of Intel Xeon E5540 CPU at 2.5 GHz and 32GB of main memory under 64-bit GNU/Linux. We considered the genomes of thirteen bacteria and four case-study eukaryotes (Table 4), all obtained from the NCBI database (<ftp://ftp.ncbi.nih.gov/genomes/>).

To test the correctness of our implementation, we compared it against the implementation of Pinho et al. [15], which we denote here by PFG. In particular, we counted the number of minimal absent words, for lengths 11, 14, 17, and 24,

$j$	Before	BeforeLCP	Factor	Minimal absent words	Tuple representation
0	11	11	A		
1	00	11	A A	AAA, BAA	$\langle A, (0, 1) \rangle, \langle B, (0, 1) \rangle$
2	11	11	A B		
3	10	11	A B A B	BABABA	$\langle B, (1, 5) \rangle$
4	01	11	A B A B B	AABABB	$\langle A, (3, 7) \rangle$
5	11	11	A B A		
6	01	11	A B B	AABB	$\langle A, (5, 7) \rangle$
7	11	11	A		
8	11	11	B		
9	01	11	B	We do not consider this row as it corresponds to the end of the word $y$	
10	10	11	B A	BBA	$\langle B, (2, 3) \rangle$
11	10	10	B A B A		
12	10	10	B A B B		
13	10	11	B A	BBA	This is a duplicate so we ignore it
14	10	11	B B	BBB	$\langle B, (6, 7) \rangle$

Table 3: Minimal absent words of word  $y = AABABABB$ ; we find seven minimal absent words  $\{AAA, AABABB, AABB, BAA, BABABA, BBA, BBB\}$

Species	Abbreviation	Genome reference
<b>Bacteria</b>		
<i>Bacillus anthracis</i> strain Ames	Ba	NC003997
<i>Bacillus subtilis</i> strain 168	Bs	NC000964
<i>Escherichia coli</i> strain K-12 substrain MG1655	Ec	NC000913
<i>Haemophilus influenzae</i> strain Rd KW20	Hi	NC000907
<i>Helicobacter pylori</i> strain 26695	Hp	NC000915
<i>Lactobacillus casei</i> strain BL23	Lc	NC010999
<i>Lactococcus lactis</i> strain I1403	Ll	NC002662
<i>Mycoplasma genitalium</i> strain G37	Mg	NC000908
<i>Staphylococcus aureus</i> strain N315	Sa	NC002745
<i>Streptococcus pneumoniae</i> strain CGSP14	Sp	NC010582
<i>Xanthomonas campestris</i> strain 8004	Xc	NC007086
<b>Eukaryotes</b>		
<i>Arabidopsis thaliana</i> (thale cress)	At	AGI release 7.2
<i>Drosophila melanogaster</i> (fruit fly)	Dm	FlyBase release 5
<i>Homo sapiens</i> (human)	Hs	build 38
<i>Mus musculus</i> (mouse)	Mm	build 38

Table 4: Species selected for this work with reference to the respective abbreviation and identification of genome sequence data by accession number for bacteria or genome assembly project for eukaryotes

in the genomes of the thirteen bacteria listed in Table 4. We considered only the  $5' \rightarrow 3'$  DNA strand. Table 5 depicts the number of minimal absent words in these sequences. We denote by  $M_{11}$ ,  $M_{14}$ ,  $M_{17}$ , and  $M_{24}$  the size of the resulting sets of minimal absent words for lengths 11, 14, 17, and 24 respectively. Identical number of minimal absent words for these lengths were also reported by PFG, suggesting that our implementation is correct.

To evaluate the efficiency of our implementation, we compared it against the corresponding performance of PFG, which is currently the fastest available im-

Species	Genome size (bp)	M <sub>11</sub>	M <sub>14</sub>	M <sub>17</sub>	M <sub>24</sub>
Ba	5,227,293	1,113,398	1,001,357	32,432	46
Bs	4,214,630	951,273	1,703,309	86,372	226
Ec	4,639,675	1,072,074	1,125,653	36,395	247
Hi	1,830,023	722,860	294,353	12,158	91
Hp	1,667,825	564,308	336,122	19,276	75
Lc	3,079,196	1,126,363	502,861	13,083	246
Ll	2,365,589	764,006	507,490	25,667	183
Mg	1,664,957	246,342	66,324	2,737	28
Sa	2,814,816	755,483	704,147	32,054	138
Sp	2,209,198	904,815	327,713	10,390	234
Xc	5,148,708	804,034	1,746,214	179,346	633

Table 5: Number of minimal absent words of lengths 11, 14, 17, and 24 in the genomes of thirteen bacteria.

plementation for computing minimal absent words. We computed all minimal absent words for each chromosome sequence of the genomes of the four eukaryotes listed in Table 4. We considered both the  $5' \rightarrow 3'$  and the  $3' \rightarrow 5'$  DNA strands. Tables 6 and 7 depict elapsed-time comparisons of MAW and PFG. MAW scales linearly and is the fastest in *all* cases. It accelerates the computations by more than a factor of 2, when the length of the sequences grows, compared to PFG. MAW also reduces the memory requirements by a factor of 5 compared to PFG. The maximum allocated memory (per task) was 6GB for MAW and 30GB for PFG.

(a) At				(b) Dm			
Chromosome	Size (bp)	MAW (s)	PFG (s)	Chromosome	Size (bp)	MAW (s)	PFG (s)
1	30,427,671	40.20	51.90	2L	23,011,544	30.01	40.85
2	19,698,289	25.86	32.94	2R	21,146,708	27.52	38.38
3	23,459,830	30.84	42.30	3L	24,543,557	32.00	45.13
4	18,585,056	24.65	31.42	3R	27,905,053	36.44	48.36
5	26,975,502	35.38	48.91	X	22,422,827	29.38	40.09

Table 6: Elapsed-time comparison of MAW and PFG for computing all minimal absent words in the genome of *Arabidopsis thaliana* and *Drosophila melanogaster*

To further evaluate the efficiency of our implementation, we compared it against the corresponding performance of PFG using synthetic data. As basic dataset we used chromosome 1 of Hs. We created five instances  $S_1$ ,  $S_2$ ,  $S_3$ ,  $S_4$ , and  $S_5$  of this sequence by randomly choosing 10%, 20%, 30%, 40%, and 50% of the positions, respectively, and randomly replacing the corresponding letters to one of the four letters of the DNA alphabet. We computed all minimal absent

(a) Hs				(b) Mm			
Chromosome	Size (bp)	MAW (s)	PFG (s)	Chromosome	Size (bp)	MAW (s)	PFG (s)
1	248,956,422	426.39	972.52	1	197,195,432	340.59	599.86
2	242,193,529	423.19	772.89	2	181,748,087	316.17	578.2
3	198,295,559	353.60	645.45	3	159,599,783	274.46	506.73
4	190,214,555	339.02	616.26	4	155,630,120	266.67	473.97
5	181,538,259	342.53	577.05	5	152,537,259	260.50	424.24
6	170,805,979	299.72	538.34	6	149,517,037	256.36	455.11
7	159,345,973	305.26	491.32	7	152,524,553	257.65	413.37
8	145,138,636	254.17	437.18	8	131,738,871	223.09	344.92
9	138,394,717	235.14	356.08	9	124,076,172	210.37	334.25
10	133,797,422	235.38	392.45	10	129,993,255	222.36	363.34
11	135,086,622	236.80	379.15	11	121,843,856	208.55	324.54
12	133,275,309	235.14	390.46	12	121,257,530	205.09	324.79
13	114,364,328	191.64	269.52	13	120,284,312	204.80	314.56
14	107,043,718	178.00	240.93	14	125,194,864	212.59	336.49
15	101,991,189	167.89	222.98	15	103,494,974	175.21	265.92
16	90,338,345	153.07	198.49	16	98,319,150	166.10	249.03
17	83,257,441	144.32	207.02	17	95,272,651	160.70	232.79
18	80,373,285	137.68	199.44	18	90,772,031	153.40	223.56
19	58,617,616	100.95	126.82	19	61,342,430	101.89	125.85
20	64,444,167	109.80	144.83	X	166,650,296	282.21	503.98
21	46,709,983	74.60	74.65	Y	91,744,698	141.79	251
22	50,818,468	70.49	73.34				
X	156,040,895	275.14	457.2				
Y	57,227,415	60.85	62.34				

Table 7: Elapsed-time comparison of MAW and PFG for computing all minimal absent words in the genome of *Homo Sapiens* and *Mus musculus*

words for each instance. We considered both the  $5' \rightarrow 3'$  and the  $3' \rightarrow 5'$  DNA strands. Table 8 depicts elapsed-time comparisons of MAW and PFG. MAW is the fastest in *all* cases.

Sequence	Size (bp)	MAW (s)	PFG (s)
S <sub>1</sub>	248,956,422	435.63	746.93
S <sub>2</sub>	248,956,422	438.52	733.69
S <sub>3</sub>	248,956,422	444.62	726.34
S <sub>4</sub>	248,956,422	444.06	743.29
S <sub>5</sub>	248,956,422	449.25	741.01

Table 8: Elapsed-time comparison of MAW and PFG for computing all minimal absent words in synthetic data

## 5 Final Remarks

We presented the first  $\mathcal{O}(n)$ -time and  $\mathcal{O}(n)$ -space algorithm for computing all minimal absent words based on the construction of suffix arrays. Experimental results show that the respective implementation outperforms existing tools.

## References

1. Claudia Acquisti, George Poste, David Curtiss, and Sudhir Kumar. Nullomers: Really a matter of natural selection? *PLoS ONE*, 2(10), 2007.
2. Djamal Belazzougui, Fabio Cunial, Juha Kärkkäinen, and Veli Mäkinen. Versatile succinct representations of the bidirectional burrows-wheeler transform. In Hans L. Bodlaender and Giuseppe F. Italiano, editors, *ESA*, volume 8125 of *Lecture Notes in Computer Science*, pages 133–144. Springer, 2013.
3. Supaporn Chairungsee and Maxime Crochemore. Using minimal absent words to build phylogeny. *Theoretical Computer Science*, 450(0):109–116, 2012.
4. Maxime Crochemore, Filippo Mignosi, and Antonio Restivo. Automata and forbidden words. *Information Processing Letters*, 67:111–117, 1998.
5. Mirjana Domazet-Lošo and Bernhard Haubold. Efficient estimation of pairwise distances between genomes. *Bioinformatics*, 25(24):3221–3227, 2009.
6. Johannes Fischer. Inducing the LCP-Array. In Frank Dehne, John Iacono, and Jrg-Rdiger Sack, editors, *Algorithms and Data Structures*, volume 6844 of *Lecture Notes in Computer Science*, pages 374–385. Springer Berlin Heidelberg, 2011.
7. Sara P. Garcia and Armando J. Pinho. Minimal Absent Words in Four Human Genome Assemblies. *PLoS ONE*, 6(12), 2011.
8. Sara P. Garcia, O J. Pinho, Joo M. O. S. Rodrigues, Carlos A. C. Bastos, and Paulo J. S. G. Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS ONE*, 6, 2011.
9. Gregory Hampikian and Tim Andersen. Absent sequences: Nullomers and primes. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany Murray, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 355–366. World Scientific, 2007.
10. Bernhard Haubold, Nora Pierstorff, Friedrich Möller, and Thomas Wiehe. Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics*, 6:123, 2005.
11. Julia Herold, Stefan Kurtz, and Robert Giegerich. Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, 9(1):1–9, 2008.
12. Udi Manber and Eugene W. Myers. Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.
13. Filippo Mignosi, Antonio Restivo, and Marinella Sciortino. Words and forbidden factors. *Theor. Comput. Sci.*, 273(1-2):99–117, 2002.
14. Ge Nong, Sen Zhang, and Wai Hong Chan. Linear suffix array construction by almost pure induced-sorting. In *Proceedings of the 2009 Data Compression Conference, DCC '09*, pages 193–202, Washington, DC, USA, 2009. IEEE Computer Society.
15. Armando J. Pinho, Paulo J. S. G. Ferreira, and Sara P. Garcia. On finding minimal absent words. *BMC Bioinformatics*, 11, 2009.
16. Zong-Da Wu, Tao Jiang, and Wu-Jie Su. Efficient computation of shortest absent words in a genomic sequence. *Information Processing Letters*, 110(14-15):596 – 601, 2010.